LG-256638-OLS-24, Virginia Polytechnic Institute and State University (Libraries)

Harnessing ETDs: Pioneering AI-Driven Innovations in Library Service

William A. Ingram & Edward A. Fox

Introduction

Virginia Tech's University Libraries, in collaboration with the Digital Library Research Lab in the Department of Computer Science, seeks IMLS support for an applied research grant that aims to significantly expand the reach and impact of academic library services through integration with Large Language Models (LLMs). This project builds on the foundational work supported by a 2019–2023 IMLS grant (LG-37-19-0078-19), which pioneered the application of machine learning to enhance libraries' capacity to provide access to knowledge buried in electronic theses and dissertations (ETDs) and other long-form scholarship. This proposal is a substantial leap forward from our previous work, shifting from enhancing accessibility to actively integrating content and LLMs in innovative, service-oriented applications for libraries. Our aim is to take advantage of modern AI techniques to enhance the depth and reach of library collections, support graduate students, and foster data-driven academic strategies for libraries and the universities they serve. This project aligns with the NLG Program's 3rd goal—improve the ability of libraries and archives to provide broad access to and use of information and collections—particularly under Objective 3.2, which supports the design and development of online library and archives services that meet user expectations for operating in an online environment.

Project Justification

Libraries have managed access to ETDs for decades, resulting in millions available online under open licenses. Despite their richness, the usage of these extensive works remains limited, due in part to their length and complexity, which makes manual exploration prohibitively time-consuming. But as these collections grow, prospects for computational analysis and utilization are burgeoning as well, paving the way for innovative library services rooted in collections as data. Our recent IMLS-funded collaborative research project centered on enhancing computational access to book-length documents, primarily ETDs. Our work culminated successfully, with a multi-terabyte corpus of over 500K ETDs (Uddin et al., 2021); novel techniques to segment (Ahuja et al., 2023), extract (Kahu et al., 2021), and summarize (Banerjee et al., 2022) sections of ETDs; and innovative user interfaces to maximize their equity and reach. Despite this success, the ongoing evolution of AI, particularly in the realm of LLMs, accentuates a pressing need for more research.

There are growing concerns about the reliability of LLMs, their propensity to "hallucinate" facts or present inaccurate information with a degree of confidence that could mislead users. Libraries, as traditional stewards of knowledge and truth, can play a pivotal role in mitigating these challenges. Our project centers on the use of library collections, specifically a curated ETD corpus, to contextualize and validate AI-generated responses. By investigating how to anchor LLMs in our rich corpus of graduate research, we aim to increase the effectiveness of libraries to serve their communities by leveraging local scholarship. We propose a three-phase research project, starting with the contextualization of AI responses using library collections, followed by a focused exploration into applications that could serve the distinct needs of graduate students, as well as the administrative and academic personnel within the university. We seek answers to the following pivotal research questions. **RQ1**: How can libraries integrate a curated ETD corpus with LLMs to enhance the accuracy and reliability of synthesized information? **RQ2**: How can integrating ETDs with LLMs facilitate the development of AI-driven tools to meet the comprehensive needs of graduate students? **RQ3**: How can the integration of ETDs with LLMs empower libraries to devise analytical tools that provide fresh perspectives on research trends and foster informed, data-driven strategic planning?

Project Work Plan

Our central hypothesis is that integrating a curated ETD corpus with an LLM can significantly enhance the precision and reliability of information synthesis, providing a foundation for the development of AI-driven tools that can enrich library services, foster holistic support for graduate students, and offer novel insights into research trends, thus facilitating informed strategic decision-making. Over three years, we will test our hypothesis by sequentially addressing our research questions.

In Year 1, focused on **RQ1**, we will develop a prototype system that integrates LLMs with our curated ETD corpus to enhance the reliability of synthesized information. Our experimental setup will be based on LLaMA 2 (Large Language Model Meta AI) (Touvron et al., 2023) and use techniques such as retrieval-augmented generation (Lewis et al., 2020) and fine-tuning to anchor the model's responses in our ETD corpus. Building on over 40 years

in information retrieval research at VT, including recent advancements using the ETD corpus, we will ensure the LLM retrieves the most pertinent and accurate scholarly information to offer a range of library services. We will develop a test bed to rigorously evaluate the performance of the system, specifically focusing on the quality and accuracy of generated results. The test bed will support a range of controlled, repeatable experiments that consider not just quantitative accuracy but also the qualitative value of the generated text. In Year 2, focused on **RQ2**, we will investigate the feasibility of using LLMs and the ETD corpus to develop AI-driven tools to meet the comprehensive needs of graduate students. We will explore the potential of AI in shaping a comprehensive support ecosystem, aiding in aspects such as literature review, data analysis, effective writing, and citation management, as well as facilitating insights into effective methodologies and experiment setups popular within specific fields. Such a system could also serve as a conduit to foster connections within the academic community, suggesting relevant conferences, workshops, and networking opportunities based on insights gleaned from the corpus. It could also offer career guidance by providing data-driven insights into evolving opportunities and trajectories in various fields, helping students align their research efforts with broader industry and academic trends. Recognizing the multifaceted nature of the "whole student," we will investigate how AI tools can adapt and possibly anticipate changing academic needs of graduate students, thus facilitating a more responsive and nuanced academic research ecosystem. The prototypes we develop will be tested by graduate students in real-life settings. In Year 3, focused on **RQ3**, we will investigate the potential of using LLMs and the ETD corpus to analyze emerging research trends, unearth underrepresented research domains, and foster data-driven decision making at an institutional level, thus helping inform strategic planning and resource allocation. To our knowledge, no other system uses the rich data available within ETDs to derive actionable insights. We will address this gap by providing a robust platform for real-time reporting and query-based analysis of the ETD corpus. This exploration seeks not only to identify emerging interdisciplinary research avenues, but also to guide essential resource allocation and foster the creation of new interdisciplinary hubs, identifying both underrepresented research areas and potential thematic redundancies. The prototypes will be evaluated at Virginia Tech, where their impact on strategic decision making and insight generation will be assessed against the needs of administrative and academic staff within the university, exploring the potential of a cohesive system that enhances both research and teaching and fosters data-driven strategic planning at the organizational level.

During each phase, we will actively seek to publish findings in peer-reviewed journals and conferences to share knowledge and foster collaboration with other researchers in the field. To help guide our project and keep us on track, we will set up an advisory board of stakeholders that will regularly meet with our team to evaluate our progress and provide necessary guidance.

Project Results

The project will yield the following deliverables: (1) A comprehensive report on the integration of LLMs and the ETD corpus, reporting the methodology, the process, and the results of our experiments. (2) Working prototypes of software that integrates the ETD corpus and LLMs. (3) A test bed for evaluating our software to fit the needs of multiple persona. (4) Research articles and presentations. (5) Workshops and training modules. (6) Open source software and datasets. We are committed to making our digital products freely available, managing them responsibly, and clearly reporting our methods and progress to IMLS for evaluation.

Our project is intended to optimize impact and foster interdisciplinary scholarship. Most of the funding is directed toward supporting a full-time Graduate Research Assistant (GRA) from CS to work in the library for three years. This investment not only advances research, but also significantly benefits student education and professional growth. The project will also provide a framework for course projects, offering hands-on learning experiences that can be directly implemented in the classroom. This aligns with our ethos of integrating students into library-based, grant-funded research, making experiential learning a central component of the project.

Budget Summary

We repectfully request $449,832 in IMLS grant funding: $206,058 in salaries (PI: 1 calendar month per year, co-PI: .75 summer months per year, 1 full-time 12 month GRA per year), $29,103 in fringe benefits, $57,574 tuition reimbursement for GRA, $10,000 for travel, and $147,097 in indirect costs.